Improving LSA-based Summarization with Anaphora Resolution

Josef Steinberger
University of West Bohemia
Univerzitni 22, Pilsen 30614,
Czech Republic
jstein@kiv.zcu.cz

Mijail A. Kabadjov
University of Essex
Wivenhoe Park, Colchester CO4 3SQ,
United Kingdom
malexa@essex.ac.uk

Massimo Poesio
University of Essex
Wivenhoe Park, Colchester CO4 3SQ,
United Kingdom
poesio@essex.ac.uk

Olivia Sanchez-Graillet
University of Essex
Wivenhoe Park, Colchester CO4 3SQ,
United Kingdom
osanch@essex.ac.uk

Abstract

We propose an approach to summarization exploiting both lexical information and the output of an automatic anaphoric resolver, and using Singular Value Decomposition (SVD) to identify the main terms. We demonstrate that adding anaphoric information results in significant performance improvements over a previously developed system, in which only lexical terms are used as the input to SVD. However, we also show that how anaphoric information is used is crucial: whereas using this information to add new terms does result in improved performance, simple substitution makes the performance worse.

1 Introduction

Many approaches to summarization can be very broadly characterized as TERM-BASED: they attempt to identify the main 'topics,' which generally are TERMS, and then to extract from the document the most important information about these terms (Hovy and Lin, 1997). These approaches can be divided again very broadly in 'lexical' approaches, among which we would include LSA-based approaches, and 'coreference-based' approaches . Lexical approaches to term-based summarization use lexical relations to identify central terms (Barzilay and Elhadad, 1997; Gong and Liu, 2002); coreference- (or anaphora-) based approaches (Baldwin and Morton, 1998; Boguraev and Kennedy, 1999; Bergler et al., 2003; Stuckardt, 2003) identify these terms by running a coreference- or anaphoric resolver over the text.[1] We are not aware, however, of any attempt to use both lexical and anaphoric information to identify the main terms. In addition, to our knowledge no authors have convincingly demonstrated that feeding anaphoric information to a summarizer significantly improves the performance of a summarizer using a standard evaluation procedure (a reference corpus and baseline, and widely accepted evaluation measures).

In this paper we compare two sentence extraction-based summarizers. Both use Latent Semantic Analysis (LSA) (Landauer, 1997) to identify the main terms of a text for summarization; however, the first system (Steinberger and Jezek, 2004), discussed in Section 2, only uses lexical information to identify the main topics, whereas the second system exploits both lexical and anaphoric information. This second system uses an existing anaphora resolution system to resolve anaphoric expressions, GUITAR (Poesio and Kabadjov, 2004); but, crucially, two different ways of using this information for summarization were tested. (Section 3.) Both summarizers were tested over the CAST corpus (Orasan et al., 2003), as discussed in Section 4, and significant improvements were observed over both the

[1] The terms 'anaphora resolution' and 'coreference resolution' have been variously defined (Stuckardt, 2003), but the latter term is generally used to refer to the coreference task as defined in MUC and ACE. We use the term 'anaphora resolution' to refer to the task of identifying successive mentions of the same discourse entity, realized via any type of noun phrase (proper noun, definite description, or pronoun), and whether such discourse entities 'refer' to objects in the world or not.

baseline CAST system and our previous LSA-based summarizer.

2 An LSA-based Summarizer Using Lexical Information Only

LSA (Landauer, 1997) is a technique for extracting the 'hidden' dimensions of the semantic representation of terms, sentences, or documents, on the basis of their contextual use. It is a very powerful technique already used for NLP applications such as information retrieval (Berry et al., 1995) and text segmentation (Choi et al., 2001) and, more recently, multi- and single-document summarization.

The approach to using LSA in text summarization we followed in this paper was proposed in (Gong and Liu, 2002). Gong and Liu propose to start by creating a term by sentences matrix $A = [A_1, A_2, \ldots, A_n]$, where each column vector A_i represents the weighted term-frequency vector of sentence i in the document under consideration. If there are a total of m terms and n sentences in the document, then we will have an $m \times n$ matrix A for the document. The next step is to apply Singular Value Decomposition (SVD) to matrix A. Given an $m \times n$ matrix A, the SVD of A is defined as:

(1) $\quad A = U\Sigma V^T$

where $U = [u_{ij}]$ is an $m \times n$ column-orthonormal matrix whose columns are called left singular vectors, $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is an $n \times n$ diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order, and $V = [v_{ij}]$ is an $n \times n$ orthonormal matrix, whose columns are called right singular vectors.

From a mathematical point of view, applying SVD to a matrix derives a mapping between the m-dimensional space spawned by the weighted term-frequency vectors and the r-dimensional singular vector space. From a NLP perspective, what the SVD does is to derive the *latent semantic structure* of the document represented by matrix A: a breakdown of the original document into r linearly-independent base vectors ('topics'). Each term and sentence from the document is jointly indexed by these 'topics'.

A unique SVD feature is that it is capable of capturing and modelling interrelationships among terms so that it can semantically cluster terms and sentences. Furthermore, as demonstrated in (Berry et al., 1995), if a word combination pattern is salient and recurring in document, this pattern will be captured and represented by one of the singular vectors. The magnitude of the corresponding singular value indicates the importance degree of this pattern within the document. Any sentences containing this word combination pattern will be projected along this singular vector, and the sentence that best represents this pattern will have the largest index value with this vector. As each particular word combination pattern describes a certain topic in the document, each singular vector can be viewed as representing a salient topic of the document, and the magnitude of its corresponding singular value represents the degree of importance of the salient topic.

The summarization method proposed by Gong and Liu (2002) should now be easy to understand. The matrix V^T describes the importance degree of each 'implicit topic' in each sentence: the summarization process simply chooses the most informative sentence for each term. In other words, the kth sentence chosen is the one with the largest index value in the kth right singular vector in matrix V^T.

The summarization method proposed by Gong and Liu has some disadvantages as well, the main of which it that it is necessary to use the same number of dimensions as is the number of sentences we want to choose for a summary. However, the higher the number of dimensions of reduced space is, the less significant topic we take into a summary. In order to remedy this problem, we (Steinberger and Jezek, 2004) proposed the following modifications to Gong and Liu's summarization method. After computing the SVD of a term by sentences matrix, we compute the length of each sentence vector in matrix V. This is to favour the index values in the matrix V that correspond to the highest singular values (the most significant topics). Formally:

(2) $\quad s_k = \sqrt{\sum_{i=1}^{r} v_{k,i}^2 \cdot \sigma_i^2},$

where s_k is the length of the vector of k'th sentence in the modified latent vector space, and its significance score for summarization too. The level of dimensionality reduction (r) is essentially learned from the data. Finally, we put into the summary the sentences with the highest values in vector s. We showed in previous work (Steinberger and Jezek, 2004) that this modification results in a significant

improvement over Gong and Liu's method.

3 Using Anaphora Resolution for Summarization

3.1 The case for anaphora resolution

Words are the most basic type of 'term' that can be used to characterize the content of a document. However, being able to identify the most important *objects* mentioned in the document clearly would lead to an improved analysis of what is important in a text, as shown by the following news article cited by Boguraev and Kennedy (1999):

(3) PRIEST IS CHARGED WITH POPE ATTACK

> *A Spanish priest* was charged here today with attempting to murder the Pope. *Juan Fernandez Krohn*, aged 32, was arrested after a man armed with a bayonet approached the Pope while he was saying prayers at Fatima on Wednesday night. According to the police, *Fernandez* told the investigators today that *he* trained for the past six months for the assault. ... If found guilty, *the Spaniard* faces a prison sentence of 15-20 years.

As Boguraev and Kennedy point out, the title of the article is an excellent summary of the content: an entity (the priest) did something to another entity (the pope). Intuitively, understanding that Fernandez and the pope are the central characters is crucial to provide a good summary of texts like these.[2] Among the clues that help us to identify such 'main characters', the fact that an entity is repeatedly mentioned is clearly important.

Purely lexical methods, including the LSA-based methods discussed in the previous section, can only capture part of the information about which entities are frequently repeated in the text. As example (3) shows, stylistic conventions forbid verbatim repetition, hence the six mentions of Fernandez in the text above contain only one lexical repetition, 'Fernandez'. The main problem are pronouns, that tend to share the least lexical similarity with the form used to express the antecedent (and anyway are usually removed by stopword lists, therefore do not get included in the SVD matrix). The form of definite descriptions (*the Spaniard*) doesn't always overlap with that of their antecedent, either, especially when the antecedent was expressed with a proper name. The form of mention which more often overlaps to a degree with previous mentions is proper nouns, and even then at least some way of dealing with acronyms is necessary (cfr. *European Union* / *E.U.*). The motivation for anaphora resolution is that it should tell us which entities are repeatedly mentioned.

In this work, we tested a mixed approach to integrate anaphoric and word information: using the output of the anaphoric resolver GUITAR to modify the SVD matrix used to determine the sentences to extract. In the rest of this section we first briefly introduce GUITAR, then discuss the two methods we tested to use its output to help summarization.

3.2 GUITAR: A General-Purpose Anaphoric Resolver

The system we used in these experiments, GUITAR (Poesio and Kabadjov, 2004), is an anaphora resolution system designed to be high precision, modular, and usable as an off-the-shelf component of a NL processing pipeline. The current version of the system includes an implementation of the MARS pronoun resolution algorithm (Mitkov, 1998) and a partial implementation of the algorithm for resolving definite descriptions proposed by Vieira and Poesio (2000). The current version of GUITAR does not include methods for resolving proper nouns.

3.2.1 Personal Pronoun Resolution

Mitkov (1998) developed a robust approach to pronoun resolution which only requires input text to be part-of-speech tagged and noun phrases to be identified. Mitkov's algorithm operates on the basis of antecedent-tracking preferences (referred to hereafter as "antecedent indicators"). The approach works as follows: the system identifies the noun phrases which precede the anaphor within a distance of 2 sentences, checks them for gender and number agreement with the anaphor, and then applies genre-specific antecedent indicators to the remaining candidates (Mitkov, 1998). The noun phrase with the highest aggregate score is proposed as antecedent.

[2] It should be noted that for many newspaper articles, indeed many non-educational texts, only a 'entity-centered' structure can be clearly identified, as opposed to a 'relation-centered' structure of the type hypothesized in Rhetorical Structures Theory (Knott et al., 2001; Poesio et al., 2004).

3.2.2 Definite Description Resolution

The Vieira / Poesio algorithm (Vieira and Poesio, 2000) attempts to classify each definite description as either direct anaphora, discourse-new, or bridging description. The first class includes definite descriptions whose head is identical to that of their antecedent, as in *a house ... the house*. Discourse-new descriptions are definite descriptions that refer to objects not already mentioned in the text and not related to any such object. Bridging descriptions are all definite descriptions whose resolution depends on knowledge of relations between objects, such as definite descriptions that refer to an object related to an entity already introduced in the discourse by a relation other than identity, as in *the flat ... the living room*. The Vieira / Poesio algorithm also attempts to identify the antecedents of anaphoric descriptions and the anchors of bridging ones. The current version of GUITAR incorporates an algorithm for resolving direct anaphora derived quite directly from Vieira / Poesio, as well as a statistical version of the methods for detecting discourse new descriptions (Poesio et al., 2005).

3.3 SVD over Lexical and Anaphoric Terms

SVD can be used to identify the 'implicit topics' or main terms of a document not only when on the basis of words, but also of coreference chains, or a mixture of both. We tested two ways of combining these two types of information.

3.3.1 The Substitution Method

The simplest way of integrating anaphoric information with the methods used in our earlier work is to use anaphora resolution simply as a preprocessing stage of the SVD input matrix creation. Firstly, all anaphoric relations are identified by the anaphoric resolver, and anaphoric chains are identified. Then a second document is produced, in which all anaphoric nominal expressions are replaced by the first element of their anaphoric chain. For example, suppose we have the text in (4).

(4) **S1:** *Australia's new conservative government* on Wednesday began selling *its tough deficit-slashing budget*, which sparked *violent protests by Aborigines, unions, students and welfare groups* even before *it* was announced.

S2: Two days of *anti-budget street protests* preceded *spending cuts* officially unveiled by *Treasurer Peter Costello*.

S3: "If *we* don't do *it* now, *Australia* is going to be in *deficit* and debt into the next century."

S4: As *the protesters* had feared, *Costello* revealed a cut to *the government's* Aboriginal welfare commission among *the hundreds of measures implemented to claw back the deficit*.

An ideal resolver would find 8 anaphoric chains:

Chain 1 *Australia - we - Australia*

Chain 2 *its new conservative government (Australia's new conservative government) - the government*

Chain 3 *its tough deficit-slashing budget (Australia's tough deficit-slashing budget) - it*

Chain 4 *violent protests by Aborigines, unions, students and welfare groups - anti-budget street protests*

Chain 5 *Aborigines, unions, students and welfare groups - the protesters*

Chain 6 *spending cuts - it - the hundreds of measures implemented to claw back the deficit*

Chain 7 *Treasurer Peter Costello - Costello*

Chain 8 *deficit - the deficit*

By replacing each element of the 8 chains above in the text in (4) with the first element of the chain, we get the text in (5).

(5) **S1:** *Australia's new conservative government* on Wednesday began selling *Australia's tough deficit-slashing budget*, which sparked *violent protests by Aborigines, unions, students and welfare groups* even before *Australia's tough deficit-slashing budget* was announced.

S2: Two days of *violent protests by Aborigines, unions, students and welfare groups* preceded *spending cuts* officially unveiled by *Treasurer Peter Costello*.

S3: "If *Australia* doesn't do *spending cuts* now, *Australia* is going to be in *deficit* and debt into the next century."

S4: As *Aborigines, unions, students and welfare groups* had feared, *Treasurer Peter Costello* revealed a cut to *Australia's new conservative government's* Aboriginal welfare commission among *the spending cuts*.

This text is then used to create the SVD input matrix, as done in the first system.

3.3.2 The Addition Method

An alternative approach is to use SVD to identify 'topics' on the basis of two types of 'terms': terms in the lexical sense (i.e., words) and terms in the sense of objects, which can be represented by anaphoric

chains. In other words, our representation of sentences would specify not only if they contain a certain word, but also if they contain a mention of a discourse entity (See Figure 1.) This matrix would then be used as input to SVD.

Figure 1: Addition method.

The chain 'terms' tie together sentences that contain the same anaphoric chain. If the terms are lexically the same (direct anaphors - like *deficit* and *the deficit*) the basic summarizer works sufficiently. However, Gong and Liu showed that the best weighting scheme is boolean (i.e., all terms have the same weight); our own previous results confirmed this. The advantage of the addition method is the opportunity to give higher weights to anaphors.

4 Evaluation

4.1 The CAST Corpus

To evaluate our system, we used the corpus of manually produced summaries created by the CAST project[3] (Orasan et al., 2003). The CAST corpus contains news articles taken from the Reuters Corpus and a few popular science texts from the British National Corpus. It contains information about the importance of the sentences (Hasler et al., 2003). Sentences are marked as **essential** or **important**. The corpus also contains annotations for linked sentences, which are not significant enough to be marked as important/essential, but which have to be considered as they contain information essential for the understanding of the content of other sentences marked as essential/important.

Four annotators were used for the annotation, three graduate students and one postgraduate. Three of the annotators were native English speakers, and the fourth had advanced knowledge of English. Unfortunately, not all of the documents were annotated by all of the annotators. To maximize the reliability of the summaries used for evaluation, we chose the documents annotated by the greatest number of the annotators; in total, our evaluation corpus contained 37 documents.

For acquiring manual summaries at specified lengths and getting the sentence scores (for relative utility evaluation) we assigned a score 3 to the sentences marked as essential, a score 2 to important sentences and a score 1 to linked sentences. The sentences with highest scores are then selected for ideal summary (at specified lenght).

4.2 Evaluation Measures

Evaluating summarization is a notoriously hard problem, for which standard measures like Precision and Recall are not very appropriate. The main problem with P&R is that human judges often disagree what are the top n% most important sentences in a document. Using P&R creates the possibility that two equally good extracts are judged very differently. Suppose that a manual summary contains sentences [1 2] from a document. Suppose also that two systems, A and B, produce summaries consisting of sentences [1 2] and [1 3], respectively. Using P&R, system A will be ranked much higher than system B. It is quite possible that sentences 2 and 3 are equally important, in which case the two systems should get the same score.

To address the problem with precision and recall we used a combination of evaluation measures. The first of these, relative utility (RU) (Radev et al., 2000) allows model summaries to consist of sentences with variable ranking. With RU, the model summary represents all sentences of the input document with confidence values for their inclusion in the summary. For example, a document with five sentences [1 2 3 4 5] is represented as [1/5 2/4 3/4 4/1 5/2]. The second number in each pair indicates the degree to which the given sentence should be part of the summary according to a human judge. This number is called the utility of the sentence. Utility depends on the input document, the summary length, and the judge. In the example, the system that selects sentences [1 2] will not get a higher score than a system that chooses sentences [1 3] given that both summaries [1 2] and [1 3] carry the same number of utility points (5+4). Given that no other combination of two sentences carries a higher utility, both systems [1 2] and [1 3] produce optimal extracts. To compute relative utility, a number of

[3]The goal of this project was to investigate to what extent Computer-Aided Summarization can help humans to produce high quality summaries with less effort.

Evaluation Method	Lexical LSA	Manual Substitution	Manual Addition
Relative Utility	0.595	0.573	**0.662**
F-score	0.420	0.410	**0.489**
Cosine Similarity	0.774	0.806	**0.823**
Main Topic Similarity	0.686	0.682	**0.747**

Table 1: Evaluation of the manual annotation improvement - summarization ratio: 15%.

Evaluation Method	Lexical LSA	Manual Substitution	Manual Addition
Relative Utility	0.645	0.662	**0.688**
F-score	0.557	0.549	**0.583**
Cosine Similarity	0.863	0.878	**0.886**
Main Topic Similarity	0.836	0.829	**0.866**

Table 2: Evaluation of the manual annotation improvement - summarization ratio: 30%.

judges, $(N \geq 1)$ are asked to assign utility scores to all n sentences in a document. The top e sentences according to utility score are then called a sentence extract of size e. We can then define the following system performance metric:

$$RU = \frac{\sum_{j=1}^{n} \delta_j \sum_{i=1}^{N} u_{ij}}{\sum_{j=1}^{n} \epsilon_j \sum_{i=1}^{N} u_{ij}}, \quad (6)$$

where u_{ij} is a utility score of sentence j from annotator i, ϵ_j is 1 for the top e sentences according to the sum of utility scores from all judges and δ_j is equal to 1 for the top e sentences extracted by the system. For details see (Radev et al., 2000).

The second measure we used is Cosine Similarity, according to the standard formula:

$$cos(X, Y) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i (x_i)^2} \cdot \sqrt{\sum_i (y_i)^2}}, \quad (7)$$

where X and Y are representations of a system summary and its reference summary based on the vector space model. The third measure is Main Topic Similarity. This is a content-based evaluation method based on measuring the cosine of the angle between first left singular vectors of a system summary's and its reference summary's SVDs. (For details see (Steinberger and Jezek, 2004).) Finally, we measured ROUGE scores, with the same settings as in the Document Understanding Conference (DUC) 2004.

4.3 How Much May Anaphora Resolution Help? An Upper Bound

We annotated all the anaphoric relations in the 37 documents in our evaluation corpus by hand using the annotation tool MMAX (Mueller and Strube, 2003). [4] Apart from measuring the performance of GUITAR over the corpus, this allowed us to establish the upper bound on the performance improvements that could be obtained by adding an anaphoric resolver to our summarizer. We tested both methods of adding the anaphoric knowledge to the summarizer discussed above. Results for the 15% and 30% ratios[5] are presented in Tables 1 and 2. The baseline is our own previously developed LSA-based summarizer without anaphoric knowledge. The result is that the substitution method did not lead to significant improvement, but the addition method did: addition could lead to an improvement in Relative Utility score from .595 to .662 for the 15% ratio, and from .645 to .688 for the 30% ratio.

4.4 Results with GUITAR

To use GUITAR, we first parsed the texts using Charniak's parser (Charniak, 2000). The output of the parser was then converted into the MAS-XML format expected by GUITAR by one of the preprocessors that come with the system. (This step includes heuristic methods for guessing agreement features.) Finally, GUITAR was ran to add anaphoric information to the files. The resulting files were then processed by the summarizer.

GUITAR achieved a precision of 64.0% and a recall of 52.7% over the 37 documents. For definite description resolution, we found a precision of

[4] We annotated personal pronouns, definite descriptions and also proper nouns, who will be handled by a future GUITAR version.

[5] We used the same summarization ratios as in CAST.

Evaluation Method	Lexical LSA	CAST	GUITAR Substitution	GUITAR Addition
Relative Utility	0.595	0.527	0.591	**0.624**
F-score	0.420	0.348	0.413	**0.445**
Cosine Similarity	0.774	0.726	0.774	**0.789**
Main Topic Similarity	0.686	0.630	0.645	**0.691**

Table 3: Evaluation of the GUITAR improvement - summarization ratio: 15%.

Evaluation Method	Lexical LSA	CAST	GUITAR Substitution	GUITAR Addition
Relative Utility	0.645	0.618	0.648	**0.663**
F-score	0.557	0.522	0.540	**0.560**
Cosine Similarity	**0.863**	0.855	0.858	0.857
Main Topic Similarity	**0.836**	0.810	0.810	0.819

Table 4: Evaluation of the GUITAR improvement - summarization ratio: 30%.

78.3% and a recall of 56.0%; for pronoun resolution, the precision was 47.8%, recall was 46.8%.

The results with the summarizer are presented in Tables 3 and 4 (relative utility, f-score, cosine, and main topic), and 5-6 (ROUGE). The contribution of the different anaphora resolution components is addressed in (Kabadjov et al., 2005). All versions of our summarizer (the baseline version without anaphora resolution and those using substitution and addition) outperformed the CAST summarizer, but we have to emphasize that CAST did not aim at producing a high-performance generic summarizer; only a system that could be easily used for didactical purposes. However, our tables also show that using GUITAR and the addition method lead to significant improvements over our baseline LSA summarizer. The improvement in Relative Utility measure was significant by t-test at 95% confidence. On the other hand, the substitution method did not lead to significant improvements, as was to be expected given that no improvement was obtained with 'perfect' anaphora resolution (see previous section).

5 Conclusion and Further Research

Our main result in this paper is to show that using anaphora resolution in summarization can lead to significant improvements, not only when 'perfect' anaphora information is available, but also when an automatic resolver is used, provided that the anaphoric resolver has reasonable performance. As far as we are aware, this is the first time that such a result has been obtained using standard evaluation measures over a reference corpus. We also showed however that the way in which anaphoric information is used matters: with our set of documents at least, substitution would not result in significant improvements even with perfect anaphoric knowledge.

Further work will include, in addition to extending the set of documents and testing the system with other collections, evaluating the improvement to be achieved by adding a proper noun resolution algorithm to GUITAR.

References

B. Baldwin and T. S. Morton. 1998. Dynamic coreference-based summarization. In *Proc. of EMNLP*. Granada, Spain.

R. Barzilay and M. Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain.

S. Bergler, R. Witte, M. Khalife, Z. Li, and F. Rudzicz. 2003. Using Knowledge-poor Coreference Resolution for Text Summarization. In *Proceedings of DUC*. Edmonton.

M. W. Berry, S. T. Dumais and G. W. O'Brien. 1995. Using Linear Algebra for Intelligent IR. In *SIAM Review*, 37(4).

B. Boguraev and C. Kennedy. 1999. Salience-based content characterization of text documents. In I. Mani and M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press. Cambridge, MA.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL*. Philadelphia.

F. Y. Y. Choi, P. Wiemer-Hastings and J. D. Moore. 2001. Latent Semantic Analysis for Text Segmentation. In *Proceedings of EMNLP*. Pittsburgh.

Y. Gong and X. Liu. 2002. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proceedings of ACM SIGIR*. New Orleans.

ROUGE measure	Baseline	CAST	GUITAR Substitution	GUITAR Addition
ROUGE-1	0.786	0.676	0.717	**0.809**
ROUGE-2	0.754	0.631	0.670	**0.781**
ROUGE-3	0.745	0.626	0.656	**0.773**
ROUGE-4	0.740	0.624	0.643	**0.767**
ROUGE-L	0.288	0.240	0.263	**0.295**
ROUGE-W-1.2	0.009	0.009	**0.011**	0.009

Table 5: ROUGE scores - summarization ratio: 15%.

ROUGE measure	Baseline	CAST	GUITAR Substitution	GUITAR Addition
ROUGE-1	**0.663**	0.602	0.641	0.660
ROUGE-2	**0.609**	0.529	0.581	0.605
ROUGE-3	0.568	0.493	0.543	**0.569**
ROUGE-4	0.535	0.465	0.511	**0.537**
ROUGE-L	0.251	0.226	**0.256**	0.245
ROUGE-W-1.2	0.021	0.014	**0.023**	0.020

Table 6: ROUGE scores - summarization ratio: 30%.

L. Hasler, C. Orasan and R. Mitkov. 2003. Building better corpora for summarization. In *Proceedings of Corpus Linguistics*. Lancaster, United Kingdom.

E. Hovy and C. Lin. 1997. Automated text summarization in SUMMARIST. In *ACL/EACL Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain.

M. A. Kabadjov, M. Poesio and J. Steinberger. 2005. Task-Based Evaluation of Anaphora Resolution: The Case of Summarization. In *RANLP Workshop "Crossing Barriers in Text Summarization Research"*. Borovets, Bulgaria.

A. Knott, J. Oberlander, M. O'Donnell, and C. Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In Sanders, T., Schilperoord, J., and Spooren, W. (eds), *Text representation: linguistic and psycholinguistic aspects*. John Benjamins.

T. K. Landauer and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. In *Psychological Review*, 104, 211-240.

R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of COLING*. Montreal.

C. Mueller and M. Strube. 2001. MMAX: A Tool for the Annotation of Multi-modal Corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Seattle.

C. Orasan, R. Mitkov and L. Hasler. 2003. CAST: a Computer-Aided Summarization Tool. In *Proceedings of EACL*. Budapest, Hungary.

M. Poesio and M. A. Kabadjov. 2004. A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation. In *Proceedings of LREC*. Lisbon, Portugal.

M. Poesio, R. Stevenson, B. Di Eugenio, and J. M. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3).

M. Poesio, M. A. Kabadjov, R. Vieira, R. Goulart, and O. Uryupina. 2005. Do discourse-new detectors help definite description resolution? In *Proceedings of IWCS*. Tilburg, The Netherlands.

D. R. Radev, H. Jing, and M. Budzikowska. 2000. Centroid-based summarization of multiple documents. In *ANLP/NAACL Workshop on Automatic Summarization*. Seattle.

J. Steinberger and K. Jezek. 2004. Text Summarization and Singular Value Decomposition. In *Proceedings of ADVIS*. Izmir, Turkey.

R. Stuckardt. 2003. Coreference-Based Summarization and Question Answering: a Case for High Precision Anaphor Resolution. In *International Symposium on Reference Resolution.* Venice, Italy.

R. Vieira and M. Poesio. 2000. An empirically-based system for processing definite descriptions. In *Computational Linguistics*, 26(4).